Parameterization of a computational physical model for glottal flow using inverse filtering and high-speed videoendoscopy

Murtola, Tiina

2018-02

Murtola , T , Alku , P , Malinen , J & Geneid , A 2018 , ' Parameterization of a computational physical model for glottal flow using inverse filtering and high-speed videoendoscopy ' , Speech Communication , vol. 96 , pp. 67-80 . https://doi.org/10.1016/j.specom.2017.11.007

http://hdl.handle.net/10138/300880
https://doi.org/10.1016/j.specom.2017.11.007

publishedVersion

Speech Communication 96 (2018) 67–80

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Parameterization of a computational physical model for glottal flow using inverse filtering and high-speed videoendoscopy

Tiina Murtola*,[a], Paavo Alku[a], Jarmo Malinen[b], Ahmed Geneid[c]

[a] Department of Signal Processing and Acoustics, Aalto University, School of Electrical Engineering, P.O. Box 12200, FI Aalto 00076, Finland
[b] Department of Mathematics and Systems Analysis, Aalto University, School of Science, P.O. Box 11100, FI Aalto 00076, Finland
[c] Department of Otorhinolaryngology and Phoniatrics, Head and Neck surgery, Helsinki University Hospital and University of Helsinki, Helsinki, Finland

ARTICLE INFO

Keywords:
Speech production
Glottal flow
Physical model
Vocal fold imaging

ABSTRACT

High-speed videoendoscopy, glottal inverse filtering, and physical modeling can be used to obtain complementary information about speech production. In this study, the three methodologies are combined to pursue a better understanding of the relationship between the glottal air flow and glottal area. Simultaneously acquired high-speed video and glottal inverse filtering data from three male and three female speakers were used. Significant correlations were found between the quasi-open and quasi-speed quotients of the glottal area (extracted from the high-speed videos) and glottal flow (estimated using glottal inverse filtering), but only the quasi-open quotient relationship could be represented as a linear model. A simple physical glottal flow model with three different glottal geometries was optimized to match the data. The results indicate that glottal flow skewing can be modeled using an inertial vocal/subglottal tract load and that estimated inertia within the glottis is sensitive to the quality of the data. Parameter optimisation also appears to favour combining the simplest glottal geometry with viscous losses and the more complex glottal geometries with entrance/exit effects in the glottis.

1. Introduction

Speech production is a complex phenomenon, and understanding it is desirable both clinically and technologically. Unfortunately, the location and function of many of the related organs, such as the vocal folds, makes direct observation challenging. One solution to this problem is offered by computational physics models that are based on the physiology of speech production, hereafter referred to as *physical models*.[1] These models can be used to carry out simulation experiments where direct measurements from human subjects would be invasive, disturb natural speech production, or be otherwise infeasible. Physical models have been shown to be potentially useful in clinical applications, for example, as aids for diagnosing voice disorders (Gómez-Vilda et al., 2007; Wurzbacher et al., 2006, 2008), assessing treatment outcomes (Švancara and Horáček, 2006; Zhang and Jiang, 2008), or providing theoretical background for therapy techniques (Titze, 2006). Physical models have also been used in the evaluation of glottal inverse filtering, a methodology to estimate the voice source (Alku et al., 2006, 2013; Guðnason et al., 2015).

One of the major challenges for physical models is the often large number of parameters which need to be estimated from scarce data. It can be argued that the simpler the model, the looser the connection is between model parameters and their physiological counterparts. Conversely, the more complicated the model, the more difficult the parameter estimation problem becomes. Physical models are compromises that, in spite of their idealizations, can be used to model natural phonation or to construct hypotheses that can be tested in natural speech. Hence, determining parameter values that best produce natural-like model output remains important. This study addresses physical modeling of the *glottis*, i.e., the orifice between the vibrating vocal folds. More specifically, the study focuses on a physical *glottal flow* model which links the periodic air flow through the glottis (i.e., the acoustical excitation of the most important category of speech signals, voiced sounds) and the transverse area of the glottis. The problem of determining parameter values for this physical model is addressed using glottal flow and glottal area signals obtained from natural vowel production.

Low-order physical glottal flow models are often used to produce the aerodynamic forces driving the vocal fold oscillations in glottis models where vocal folds are modeled using lumped elements. For this purpose, an expression linking the air flow to the time-varying glottal

* Corresponding author.
E-mail address: tiina.murtola@aalto.fi (T. Murtola).

[1] Computational physical models would be a more accurate term but the word computational has been left out to avoid confusion with purely computational models, such as deep neural networks.

https://doi.org/10.1016/j.specom.2017.11.007
Received 15 May 2017; Received in revised form 10 November 2017; Accepted 10 November 2017
Available online 11 November 2017

area is needed. The simplest physical flow models are based on stationary Bernoulli flow from the subglottal space to the point of minimal glottal opening and the assumption of atmospheric pressure downstream from that point (e.g., Steinecke and Herzel, 1995). In such models, glottal flow is considered to be proportional to the minimal glottal opening, and hence they are not able to capture *skewing*, a phenomenon measured from the production of natural speech (Childers et al., 1985; Hertegård and Gauffin, 1995), indicating that the flow peak in a glottal cycle is delayed with respect to the corresponding area.

The influence of the vocal tract (and to a lesser extent, the subglottal tract) is one of the major contributors to glottal flow skewing. The vocal tract load can be represented either as a single lumped impedance (Rothenberg, 1981; Titze, 1988; Aalto, 2009), resulting in an ordinary differential equation describing the flow–area relationship, or as pressure variables whose values are calculated, for example, using a wave-reflection or a transmission line model of the vocal tract (e.g., Ishizaka and Flanagan, 1972; Titze, 1984; Story and Titze, 1995; Lous et al., 1998). The approach using a single lumped impedance cannot represent the load to a great degree of accuracy, whereas load pressures calculated using more detailed resonator models can capture, e.g., formant information. On the other hand, resonator models typically require an assumed geometry since there are only a few datasets where the vocal tract (let alone the subglottal tract) geometry is measured simultaneously with other speech production related signals, such as sound pressure, electroglottogram (EGG) or glottal area from any laryngeal imaging method. (For an example of such a dataset, see Aalto et al., 2014.)

The effect of inertia of the air within the glottis is often considered to be negligible compared to other factors, and hence it is not included in most physical glottal flow models. Some of the few exceptions are the studies by Ishizaka and Flanagan (1972), Fant (1960), Pelorson et al. (1994) and Elie and Laprie (2016), who take different approaches to include the glottal inertia. The inertia formula used by Pelorson et al. (1994) and Elie and Laprie (2016) is appealing, as it can be combined with a lumped-impedance airway model to yield a flow–area relationship where the skewing of the flow is easily controlled, as shown later.

Since van den Berg et al. (1957) carried out model experiments investigating what they called frictional and turbulence losses in the larynx, the inclusion of glottal losses in flow models has varied widely. The turbulence term of van den Berg et al. (1957) has since evolved into a general transglottal pressure change term encompassing *vena contracta* effects at the entrance of the glottis and pressure recovery at the exit (e.g., Ishizaka and Flanagan, 1972; Titze, 1984; Story and Titze, 1995; Lucero, 1996), and sometimes even the effects of viscous friction (Titze, 1988). The model experiments done by Fulcher et al. (2011) showed that this entrance/exit effect is not generally negligible and has a broader range of coefficient values than used by van den Berg et al. (1957). It is worth noting, however, that the entrance/exit effect correction to the simple Bernoulli flow model only affects the constant of proportionality in the linear flow–area relationship, not the skewing.

Viscous friction losses in the glottis have been included in addition to or instead of the transglottal pressure coefficient by Ishizaka and Flanagan (1972), Pelorson et al. (1994), and Lucero (1996). The model experiments carried out by Fulcher et al. (2013) indicated a range of glottal dimensions where the Poiseuille effect is a reasonable estimation of viscous losses. They also suggested an improved power law to represent these losses in a wider range of glottal dimensions. Viscous losses serve to break the linear flow–area relationship when the vocal folds are nearly closed.

Physical models can also include other mechanisms which contribute to skewing. Cranen and Boves (1985a), for example, showed that the model of Ishizaka and Flanagan (1972) contains a load-independent skewing mechanism arising from the phase difference between the inferior and superior vocal fold edges, but this asymmetry factor causes relatively small skewing. Such model-specific mechanisms may increase the range of skewing a physical model is capable of producing, but they can be challenging to validate experimentally or implement in other models.

Physical glottal flow models can be used to compute the glottal air flow for given vocal fold displacements. These displacements can either be produced by a vocal fold oscillation model, or they can be measured from natural speech, e.g., by using high-speed videoendoscopy (HSV), which captures the movement of the vocal folds during phonation. Frame rates above 2 kHz make it possible to obtain several images of the vocal folds per each glottal cycle. In contrast to other imaging methods of the larynx (i.e., laryngeal stroboscopy, kymography, and photoglottography), HSV does not rely on quasi-steady vocal fold oscillations, selecting a single horizontal line to represent the entire vocal folds, or interpreting a possibly noisy light intensity signal. Early HSV studies of vocal fold oscillations used high-speed films, but these have since been replaced by digital imaging methods. Imaging can be done using either a flexible endoscope inserted through the nose or a rigid endoscope inserted orally. Although a rigid endoscope (as is used in this study) places limitations on the articulation task that can be performed, it also provides a higher spatial resolution for the entire vocal folds than a flexible endoscope.

Direct measurement of the glottal flow, to which physical models can be compared, is challenging, and only a few studies (Cranen and Boves, 1985a,b, 1988) have investigated glottal flow measurements from the natural production of speech. Glottal flow can, however, be estimated indirectly using glottal inverse filtering (GIF). GIF is a computational inversion methodology to estimate the glottal flow from an input signal (either a speech pressure waveform recorded outside the lips or an oral flow) based on the source-filter theory of speech production. In the GIF analysis, a filter representing the vocal tract is first constructed from the recorded input signal. The effect of the vocal tract is then removed by inverse filtering the input through the vocal tract model, thus leaving the source signal (i.e., the glottal flow or its first time derivative). Most of the developed GIF methods use a free-field speech pressure signal as the input (e.g., Wong et al., 1979; Alku, 1992) since its alternative (i.e., using the oral flow as the input) calls for using a specially constructed pneumotachograph mask (the so-called Rothenberg mask) (Rothenberg, 1973). The output of GIF, the estimated glottal flow or its time derivative, is in a form of a time-domain waveform. However, the amplitude scale of the estimated flow, including its DC component, is arbitrary unless the GIF analysis is conducted with a properly calibrated flow mask. The mask, however, suffers from drawbacks, such as distortions at high frequencies and difficulty combining other measurements, e.g., HSV with it.

Although the three speech investigation methodologies (physical modeling, HSV, and GIF) aim to extract complementary information about speech production, they have mainly been used separately. Yet some studies have used these methods pairwise. Physical models and *in vivo* HSV, for example, have been combined successfully. Lumped-element vocal fold model movements have been matched to displacements obtained by HSV in a number of studies (e.g., Eysholdt et al., 2003; Schwarz et al., 2006; Wurzbacher et al., 2008; Qin et al., 2009). In addition, similar matching has been done using videokymographs extracted from HSV data (Mergell et al., 2000; Döllinger et al., 2002; Wurzbacher et al., 2006; Mehta et al., 2011), with a loss of some information available about the rest of the vocal folds. Models matched to imaging data in this manner are capable of reproducing the oscillations to varying degrees, but HSV does not provide any direct information about aerodynamic load forces on the vocal folds. Hence, the physical models need to rely on an assumed expression for the force. The above studies use the model of Steinecke and Herzel (1995), where glottal air flow is assumed to be directly proportional to the area between the vocal folds with the exception of Mehta et al. (2011), who include interaction with subglottal and supraglottal tracts in their intraglottal pressures.

GIF and physical models have been used together in two ways. Physical models have been used by Alku et al. (2006, 2013) and

Guðnason et al. (2015) to produce signal pairs of glottal air flow and air pressure at the mouth opening which were used to evaluate GIF methods. In most other such studies, however, parametric yet non-physical models of the glottal flow (or its first time derivative), such as the Liljencrants-Fant model (Fant et al., 1985), are used instead, as their non-physicality is balanced by ease of use (Drugman et al., 2012; Airaksinen et al., 2014). In contrast to this approach of using physical models in evaluation of GIF, Drioli (2005) and Gómez-Vilda et al. (2007) used GIF to construct targets to which they matched the parameters of their physical model.

GIF and HSV (digital or film-based) have been used to observe the relationship between glottal area and flow signals (e.g., Berouti et al., 1977; Krishnamurthy and Childers, 1981; Granqvist et al., 2003; Pulakka, 2005). The two more recent studies, in particular, provided experimental evidence on skewing of the glottal air flow relative to the glottal area, but their results also illustrated that this relationship is complicated. The results of Granqvist et al. (2003) were based on only two test subjects, both experienced singers. As their measurement setup included an air flow mask, a rigid endoscope for HSV, and a dynamic articulation task, obtaining a large dataset by their methodology would be impractical. The dataset of Pulakka (2005) contained three test subjects, and it was obtained using a free-field microphone combined with a rigid endoscope for HSV, making this setup more easily scalable.

To our knowledge, the only studies combining a physical glottal flow model with both GIF and any vocal fold imaging method are Hertegård and Gauffin (1995) and Drioli and Foresti (2015a, 2015b). Hertegård and Gauffin (1995) used laryngeal stroboscopy with a flexible fibroscope, a Rothenberg mask based GIF, and a physical model based on entrance/exit effects to compare measurements of maximum and minimum glottal area with areas predicted by the model based on the measured flow. Drioli and Foresti (2015a) used HSV and matched a one-mass lumped-element model (including a crude flow model) to the fundamental frequency, open duration, and closed duration extracted from videokymographs (from HSV data) and glottal flows obtained by inverse filtering using conventional linear prediction analysis. Drioli and Foresti (2015b) updated the method of Drioli and Foresti (2015a) to include matching the model output to the entire glottal area and flow signals, and they added machine learning based refinement terms to both their glottal area and flow models. The refined model produced output signals that matched the measurements better, but the question of how the results could be used in interpretation or further development of physiologically motivated models remains open.

The aim of this investigation is to compute parameters of a physical glottal flow model from natural speech. In order to accomplish this, the three above-described methods are combined: HSV is used to image the vocal fold movements, GIF is carried out using a state-of-the-art method to estimate the glottal air flow from the free-field microphone signal (recorded simultaneously with HSV), and the physical glottal flow model is compared to the data obtained by the HSV and GIF. To avoid overfitting, the selection of the physical model components is guided by the limited sophistication of the HSV and GIF data. The goals of this study are (i) to parameterize the selected physical model to match natural vowel utterances, (ii) to identify, based on the parameterization, which model elements are most important in matching natural speech, and (iii) to describe quantitatively the relationship between the glottal flow and area. Although this study does not aim to produce a stand-alone physical model with direct clinical applications, it brings existing methodologies together and hence aids future development and validation of such models.

In order to bring HSV, GIF, and physical modeling into a common framework where they can be taken advantage of, the following approach is taken. First, the physical model is introduced in Section 2. In particular, the equations describing the relationship between glottal air flow and the position of the vocal folds are written out. In Section 3, the use of HSV and GIF to obtain the corresponding signals from natural speech is described. The methods of parameterizing the model to match the measurements are then detailed in Section 4 before describing and discussing the results.

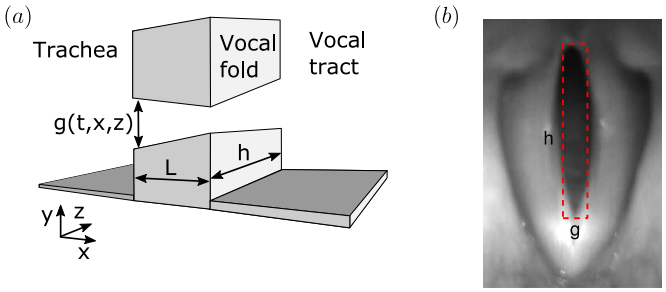

Fig. 1. Coronal view of model glottal geometry (a) showing geometry parameters glottal gap $g = g(t, x, z)$, vocal fold length h, and vocal fold thickness L. Transverse view of a rectangular glottal opening geometry superimposed on a high-speed image of the vocal folds (b).

2. A physical model of glottal flow

The proposed physical model is an idealized relationship between the glottal flow and the gap between the vocal folds. The fundamental assumptions for the model are incompressibility of the fluid and Newton's second law of motion. In addition, assumptions are made about the geometry of the glottis, as detailed below.

2.1. Simplified glottal geometries

Low-order models of the glottis require an easily computable glottal geometry; and for the model considered here, three different geometries of the space between the vocal folds are considered. The geometries are determined by three parameters as shown in Fig. 1 (a): L is the vocal fold thickness in the superior-inferior direction, h is the vocal fold length in the anterior-posterior direction, and $g = g(t, x, z)$ is the lateral distance between the vocal folds, hereafter referred to as the *glottal gap*. The origin is taken to lie at inferior and anterior corner of the vocal folds. The cross sectional area between the vocal folds is denoted $A_0(t)$ at $x = 0$ and $A_L(t)$ at $x = L$, and the minimum glottal area, which corresponds to the HSV data, is $A(t) = \min\{A_0(t), A_L(t)\}$.

The three geometries differ in g and consequently in $A_0(t)$ and $A_L(t)$:

(i) *Constant gap geometry:* $g = g(t)$, i.e., the gap is spatially constant, and hence the space between the vocal folds is a rectangular cuboid. The minimum glottal area is $A(t) = A_0(t) = A_L(t) = hg(t)$.

(ii) *Linear gap geometry:* $g = g(t, x) = g(t, 0) + \frac{x}{L}(g(t, L) - g(t, 0))$, i.e., a phase difference is allowed between the inferior and superior vocal fold edges but transverse glottal openings at all x remain rectangular. For this geometry $A_0(t) = hg(t, 0)$ and $A_L(t) = hg(t, L)$.

(iii) *Planar gap geometry:* $g = g(t, x, z) = g(t, 0, 0) + \frac{x}{L}(g(t, L, 0) - g(t, 0, 0)) + \frac{z}{h}(g(t, 0, h) - g(t, 0, 0))$, i.e., glottal gap is allowed to vary also in the anterior-posterior direction. It is further assumed that either $g(t, 0, 0) = 0$ or $g(t, L, 0) = 0$, i.e., either the superior or the inferior edge of the vocal folds form an isosceles triangle of height h and base $g(t, 0, h)$ or $g(t, L, h)$. The expressions for $A_0(t)$ and $A_L(t)$ depend on whether the glottis is converging or diverging.

The higher complexity of the linear and planar gap geometries improves their match with physiology, but their parameters are less easily estimated from high-speed images (Fig. 1 (b)). It is expected that the quality of the fit of each of the geometries to the high-speed images depends, however, on phonation type and fundamental frequency as well as on individual variations, such as anatomic factors.

2.2. The full flow model

When the vocal folds are open, the pressure balance (equivalent to Kirchoff's voltage law) from the subglottal space to the atmosphere can be written as

$$p_s = \Delta p_{a,iner} + \Delta p_{a,loss} + \Delta p_{g,iner} + \Delta p_{g,loss}, \tag{1}$$

where p_s is the subglottal stagnation pressure above the atmospheric pressure, $\Delta p_{g,iner}$ and $\Delta p_{g,loss}$ are pressure changes due inertia and non-recoverable losses within the glottis, respectively, and $\Delta p_{a,iner}$ and $\Delta p_{a,loss}$ are pressure changes due to inertia and non-recoverable losses in the airways. For this model, p_s can be taken to represent the driving pressure at any point inferior to the glottis as long as the airways in $\Delta p_{a,iner}$ and $\Delta p_{a,loss}$ comprise all parts of vocal and subglottal tracts superior to this point. For this article, p_s is assumed to be constant.

The pressure changes can be written in terms of the glottal volume velocity $U(t)$ and the glottal areas $A_0(t)$ and $A_L(t)$ giving

$$p_s = \overbrace{C_a \frac{dU(t)}{dt}}^{\Delta p_{a,iner}} + \overbrace{C_b U(t)}^{\Delta p_{a,loss}} + \overbrace{C_g \frac{d}{dt}\left(\frac{U(t)}{f_g(A_0(t), A_L(t))}\right)}^{\Delta p_{g,iner}} + \overbrace{C_v \frac{U(t)}{f_v(A_0(t), A_L(t))} + C_t \frac{U(t)^2}{A_L(t)^2}}^{\Delta p_{g,loss}}, \quad \text{for} \quad g(t) > 0, \tag{2}$$

where the parameters C_a and C_b represent airway inertance and losses, respectively, C_g glottal inertia, C_v viscous losses in the glottis, and C_t other transglottal pressure losses. All these *C parameters* are summarized in Table 1. The functions f_g and f_v depend on the chosen glottal geometry as listed in Table 2. Note that Eq. (2) is only valid when the glottis is open (i.e., $g(t) > 0$); otherwise $U(t) = 0$. Eq. (2) is henceforth referred to as the *full model*.

2.2.1. Inertial terms: lossless model

Consider Eq. (2) in the case of negligible non-recoverable glottal losses with only the inertive terms remaining

$$p_s = C_a \frac{dU(t)}{dt} + C_g \frac{d}{dt}\left(\frac{U(t)}{f_g(A_0(t), A_L(t))}\right), \quad \text{for} \quad g(t) > 0. \tag{3}$$

This equation can be derived from the unsteady Bernoulli equation between the two stagnation pressures, p_s and the atmospheric $p_{atm} = 0$, and the latter term on the right hand side arises from the velocity potential between the vocal folds. The inertance of the air column in the stationary, tubular vocal and subglottal tracts can be calculated from

$$C_a = C_{VT} + C_{SGT} = \rho \int_0^{L_{VT}} \frac{ds}{A_{VT}(s)} + \rho \int_0^{L_{SGT}} \frac{ds}{A_{SGT}(s)}, \tag{4}$$

where ρ is the density of air, $A_{VT}(s)$ is the vocal tract area at distance $s \in [0, L_{VT}]$ from the glottis, and $A_{SGT}(s)$, for $s \in [0, L_{SGT}]$, is the same for the subglottal tract. Accurate estimation of C_{VT} and C_{SGT} requires geometric data of the vocal tract and subglottal tract, but from the point of view of this glottal flow model, these appear as the lumped inertance C_a.

Inertial effects in the glottis need to be treated separately due to the moving vocal folds. Consideration of the time derivative of the velocity potential over the vocal folds (following Pelorson et al., 1994) yields glottal inertia parameter

$$C_g = \rho L_i, \tag{5}$$

where L_i is the *inertial thickness* of the vocal folds. In the constant gap geometry, L_i represents an equivalent vocal fold thickness whereas in the linear and planar gap geometries, it is a more precise description of real vocal fold dimensions.

A salient feature of the lossless case (Eq. (3)) is that it can be explicitly integrated to solve $U(t)$. This is of particular interest in the parameter estimation process to be describe in Section 4. For the integration, it is assumed (without loss of generality) that the glottal opening instant is $t_o = 0$, $U(0) = 0$, and $f_g(A_0(0), A_L(0)) = \epsilon$, where ϵ is small. Integration from t_o to a time instant t in the same pulse yields

$$U(t) = \frac{p_s t}{C_a + \frac{C_g}{f_g(A_0(t), A_L(t))}}, \quad t \in [0, t_c], \tag{6}$$

where t_c is the time instant of closure. Eq. (6) is henceforth called the *lossless model*, and the degree of flow skewing it produces is determined by the relative magnitudes of C_a and C_g as well as by the phase difference of the vocal folds. If phase difference is not large, i.e., $|A_L(t) - A_0(t)| < A_0(t)$ during most of the glottal pulse, then the lossless model is straightforward to interpret: If glottal inertia dominates, the flow is in phase with minimum glottal area $A(t)$, whereas if the airway inertia is dominant, the flow is skewed to the right.

2.2.2. Losses: viscosity, and entrance/exit effects

The third term in Eq. (2) represents viscous losses near the walls of the flow channel in the glottis. For the constant and linear gap geometries, an expression for the viscous losses can be obtained by integration of the Poiseuille formula in rectangular gaps over x. When the glottal gap varies in the z-direction as well, exact solution for these losses is difficult to obtain (see Sparrow (1962) for an approximation in ducts with isosceles triangular cross sections). Instead, integration of the Poiseuille formula over both x and z is used for the planar geometry as well although this is known to underestimate the losses at the narrow end of the glottis.[2] In all geometries, the viscous loss parameter is defined as

$$C_v = 6\mu L_v h^2, \tag{7}$$

where L_v is here referred to as *viscous thickness* and μ is the dynamic viscosity of air.

The last term in Eq. (2) accounts, in general, for the difference in the pressure drop at the glottal entrance and its recovery at the exit, seen as an extra pressure loss between the two stagnation pressures, p_s and $p_{atm} = 0$. This non-recoverable loss is, by assumption, proportional to

Table 1
C parameters, and the corresponding free physical parameters and β parameters.

C parameter	Explanation	Free physical parameters	β parameter
C_a	airway inertance		β_1
C_b	airway losses		$\widetilde{\beta}_1$
C_g	glottal inertance	L_i	β_2
C_v	viscous losses in glottis	L_v, h	β_3
C_t	entrance/exit effects in glottis	k_t	β_4

Table 2
Functions relating the glottal flow to glottal areas $A_0(t)$ and $A_L(t)$ in Eq. (2).

Gap geometry	Gap orientation	f_g	f_v
Constant	$A_0(t) = A_L(t)$	$A_0(t)$	$\frac{A_0(t)^3}{2}$
Linear	$A_0(t) \neq A_L(t)$	$\frac{A_L(t) - A_0(t)}{\ln(A_L(t)/A_0(t))}$	$\frac{A_L(t)^2 A_0(t)^2}{A_0(t) + A_L(t)}$
	$A_0(t) = A_L(t)$	$A_0(t)$	$\frac{A_0(t)^3}{2}$
Planar	$A_0(t) > A_L(t)$	$\frac{A_L(t) - A_0(t)}{\ln(A_L(t)/A_0(t))}$	$\frac{A_L(t)^2 (A_0(t) - A_L(t))}{\ln(2A_0(t)^2/(A_0(t)^2 + A_L(t)^2))}$
	$A_0(t) = A_L(t)$	$A_0(t)$	$A_0(t)^3$
	$A_0(t) < A_L(t)$	$\frac{A_L(t) - A_0(t)}{\ln(A_L(t)/A_0(t))}$	$\frac{A_0(t)^2 (A_L(t) - A_0(t))}{\ln(2A_L(t)^2/(A_0(t)^2 + A_L(t)^2))}$

[2] In an equilateral triangular flow channel, where the exact solution of the loss can be given, the integrated estimate is approximately 1/3 of the exact solution.

the kinetic energy density of the flow between the vocal folds, and hence can be thought of as an empirical correction to the lossless (steady) Bernoulli principle. Exit effects tend to be larger than entrance effects (Fulcher et al., 2011), and hence only the exit effects are included in this pressure loss term. The transglottal pressure loss parameter can be defined as

$$C_t = \frac{1}{2} k_t \rho, \tag{8}$$

where k_t is a constant.

3. Data, image processing, and glottal inverse filtering

3.1. Data collection and selection

The HSV data used in this investigation is a part of a larger multichannel (HSV, speech, EGG) dataset that was recently collected by the authors of this study for speech research purposes. The data were obtained from five male and five female speakers, each producing a vowel sound using normal (i.e., modal) and breathy phonation at low, medium, and high pitch, resulting in 60 samples, 200 ms each. No fixed targets for pitch and degree of breathiness were used as these would have made already challenging tasks more difficult and lead to increased number of repetitions. The speakers changed their phonation to produce six perceptually different vowel sounds, and the production of the utterances was monitored by an experienced experimenter. The speakers were asked to repeat the task if a sufficiently large difference was not observed. The measurements took typically 2–3 h per speaker depending on their experience, tolerance for the endoscope and heat from it, as well as anatomical and technical factors.

The HSV data collection method is shown schematically in Fig. 2. The measurements were done using the KayPentax Color High-Speed Video System (model 9710) with spatial resolution of 512 × 512 pixels and temporal resolution of 2000 frames/sec. EGG was acquired with a Glottal Enterprises electroglottograph (EG2-PCX2). A DPA omnidirectional headset microphone (model 4065-BL) was set 6.5 cm from the centre of the speaker's mouth. The microphone and EGG signals were recorded using a MOTU UltraLite-mk3 Hybrid audio interface connected to a MacBook Pro running OS X (v. 10.9.5) and AudioDesk 4. To enable synchronization of the audio signals with the video, a synchronization signal comprising binary frequency-shift keyed code at the beginning of each second was used. This signal was played in AudioDesk simultaneously with the recording and directed from the audio interface to the high-speed unit's audio capture module, as well as looped back to the audio interface as an input.

The microphone and EGG signals were high-pass filtered (cut-off frequency 60 Hz, linear phase) and synchronized to the high-speed video by aligning the synchronization signals and shifting them to account for various delays. The delays, including propagation delays and internal delays within and between the measurements systems, were estimated to be approximately 1.6 ms for males and 1.5 ms for females. After the completion of this alignment, the maximum error in the synchronization of the EGG signal to the video is ± 0.5 ms (one frame). In addition, the alignment of the microphone signal to the EGG can have an error of at most ± 0.08 ms due to the estimation of the propagation delays.

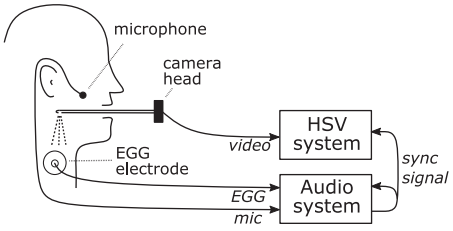

Fig. 2. A rigid endoscope connected to the HSV system is used to acquire videos of vocal fold movements, while EGG and microphone signals are recorded simultaneously. Synchronization is done by recording a custom synchronization signal with the video, EGG, and microphone signals.

Table 3
Data after selection.

Male		
Sample	f_o (Hz)	Database reference
m01	297	M02-normal-high
m02	118	M02-normal-medium
m03	173	M03-normal-high
m04	130	M03-normal-medium
m05	178	M04-normal-high
m06	114	M04-normal-low
Female		
Sample	f_o (Hz)	Database reference
f01	293	F04-normal-medium
f02	186	F04-normal-low
f03	281	F05-normal-medium
f04	185	F05-normal-low
f05	179	F03-normal-low

For this article, only the normal phonation data from the full dataset are used, and samples where the vocal folds remain partially open throughout the glottal cycle were left out. This exclusion process makes interpretation of the results of model fitting easier by justifying the assumption that there is no air flow through the glottis during the closed phase. Samples where the vocal folds are not completely visible were also removed. After the exclusions, six samples from males and five from females remained, as listed in Table 3. The remaining data is balanced in gender and represents a wide range of fundamental frequencies (114–297 Hz) and ages (25–61 years).

3.2. Glottal area extraction

Glottal area function $A_p[m]$ (in pixels) was extracted from each frame $m = 1, \ldots, 400$ of the high-speed videos. Only the red channel of the red-green-blue (RGB) images was used, and extraction was carried out using the adapted seeded region growing method developed by Lohscheller et al. (2007). This method proved to be robust against variations in image quality but suffered from periodically changing lighting conditions caused by light reflection when the vocal folds were closed. Hence, the boundaries of the extracted areas were inspected manually and, where necessary, fixed.

The pixel size is constant within each individual video recording but it varies from take to take. To convert pixel data to meter-based units, the vocal fold length in pixels h_p was estimated from an image where the glottis was fully open. This length was then assumed to correspond to a vocal fold length h (in meters), giving the glottal area function in m^2 as $A[m] = A_p[m](h/h_p)^2$.

In order to obtain the inferior and superior cross sectional areas ($A_0[m]$ and $A_L[m]$, respectively) from $A[m]$, it was assumed that $A_L(t) = A_0(t - \tau)$, where τ is the vocal fold phase delay. This delay cannot be estimated directly from the data. Instead, it is obtained by considering its consequences in the physical model (see Section 4). $A_L[m]$ was computed by shifting the closing phase of each area pulse forward in time by $\tau/2$ and resampling at the original time instants using spline interpolation. $A_0[m]$ was computed similarly but with a backward shift of $\tau/2$ in time.

As a compromise between the different sampling rates in HSV and microphone recordings, the area and gap signals were upsampled to 10 kHz using MATLAB's inbuilt function `resample`. The fluctuations introduced by this process in the closed phase of the glottal cycle were removed by forcing the signals to be zero when $A[m]$ was zero as well

as anywhere outside this interval where the resampled area signal was negative. After upsampling, area extracted from HSV and the minimum of the upsampled $A_0[m]$ and $A_L[m]$ signals may differ at the peak of the area pulses but the difference is at most 3%. Hereafter, the glottal area A refers to the upsampled $A[m]$.

3.3. Glottal inverse filtering

The microphone and EGG signals were downsampled to 10 kHz, and the microphone signals were inverse filtered using Aalto Aparat (Alku et al., 2017). Aalto Aparat is a semi-automatic GIF tool that allows the user to adjust the key parameters of GIF. The microphone signal is inverse filtered by the tool to produce both the estimated glottal flow and its first time derivative as outputs. Aalto Aparat enables estimation of the glottal flow with two GIF methods: iterative adaptive inverse filtering (Alku, 1992) and quasi-closed phase analysis (Airaksinen et al., 2014). The latter was used in the current study because it has been shown to be the most accurate GIF method in comparison with four other algorithms (Airaksinen et al., 2014). The EGG signals were used to support the inverse filtering process by visually checking that flow onset and offset aligned roughly with glottal opening and closure. Ten consecutive glottal cycles were analysed, and these were taken from the middle of the glottal flow estimated by GIF and glottal area signals.

The full flow model is not scale invariant, and hence the magnitude and the DC component of $U(t)$ matter. Unfortunately, as mentioned in the Introduction, the absolute scale of the glottal flow cannot be determined if GIF is computed from the free-field microphone signal as in the current study. Instead, the additional parameters $U_{\min}$ and $U_{\max}$ must be introduced in order to obtain a suitably scaled glottal flow for comparison with Eq. (2). The glottal flow estimate $\widetilde{U}_{IF}$ obtained by GIF is first normalized

$$\widehat{U}_{IF}[n] = \frac{\widetilde{U}_{IF}[n] - \min_n \widetilde{U}_{IF}[n]}{\max_n \widetilde{U}_{IF}[n] - \min_n \widetilde{U}_{IF}[n]}, \tag{9}$$

where n is the discrete time variable, and then scaled

$$U_{IF}[n] = s_1 \widehat{U}_{IF}[n] + s_2, \quad \text{where}$$
$$s_1 = U_{max} - U_{min} \quad \text{and} \quad s_2 = U_{min}. \tag{10}$$

As only normal phonation with full glottal closure is considered, the amplitude scaling is further simplified by assuming $U_{\min} = 0$. This reduces Eq. (10) to

$$U_{IF}[n] = U_{max} \widehat{U}_{IF}[n]. \tag{11}$$

It is worth noting that despite the normalisation, some pulses in U_{IF} may exhibit non-zero flow during the closed phase. This is a known phenomenon in GIF (see Wong et al., 1979; Alku, 2011) which is mainly due to imperfect cancellation of formants but also due to using simplified assumptions, such as time-invariance and linearity, in modelling of the human speech production mechanism in GIF.

4. Parameterization of the physical model

In order to match the HSV and GIF data of natural speech using a physical model, the parameters to the model must be optimized. In this investigation, the optimization is done using the glottal area from HSV as an input to the physical model and the glottal flow estimated by GIF as the target, as shown in Fig. 3.

The parameters that need to be optimized can be identified from Eq. (2): subglottal pressure p_s and the five C parameters. In addition, the vocal fold delay τ required in linear and planar gap geometries needs to be optimized. The C parameters can, in turn, be expressed using six free parameters as summarized in Table 1 (see also Eqs. (4), (5), (7) and (8)), but the vocal fold length h is treated as a known constant (see Table 4). There is redundancy in this parameterization, however, and by using the re-parameterization

$$\beta_1 = \frac{C_a}{p_s},\ \widetilde{\beta}_1 = \frac{C_b}{p_s},\ \beta_2 = \frac{C_i}{p_s}\cdot 10^4,\ \beta_3 = \frac{C_v}{p_s}\cdot 10^{10},\ \beta_4 = \frac{C_t}{p_s} \tag{12}$$

Eq. (2) can be written as

$$1 = \beta_1 \frac{dU(t)}{dt} + \widetilde{\beta}_1 U(t) + \frac{\beta_2}{10^4}\frac{d}{dt}\left(\frac{U(t)}{f_g(t)}\right) + \frac{\beta_3}{10^{10}}\frac{U(t)}{f_v(t)} + \beta_4 \frac{U(t)^2}{A_L(t)^2}, \tag{13}$$

where a shortened notation $f_g(t) = f_g(A_0(t), A_L(t))$ and similarly for $f_v(t)$ is used for clarity. These β *parameters* in Eq. (13) are the parameters to be numerically optimized, and their correspondence to the C parameters is summarized in Table 1. The scaling of β_2 and β_3 is done in order to avoid numerical problems caused by the parameters values differing by several orders of magnitude in the optimization.

To find the optimal β parameter values, a two-step optimization process is introduced, as shown in Fig. 3. The two steps make it possible to use different parts of the glottal cycle for determining the various parameter values. The delay τ is optimised separately, as detailed further below, since it enters into the data processing stage rather than the model simulation stage, and hence joint optimization would be challenging. For the two β optimization steps, τ is treated as an already known constant.

In Step 1 (Fig. 3 (a)), the parameters of the lossless model are optimized using a non-linear least squares (NLLS) method. The dependent variable $\mathbf{y}_a$ and the model function $\mathbf{f}_a$ are defined element-wise as

$$y_{a,n} = U_{IF}[n] \quad \text{and} \quad f_{a,n}(f_g[n], \beta_1, \beta_2) = \frac{t_n}{\beta_1 + \frac{\beta_2}{f_g[n]}}, \tag{14}$$

where t_n is the time elapsed since the last glottal opening instant.

Step 1 makes use of a weight function $\mathbf{w}_a$ (Fig. 4) which emphasizes data in the parts of the glottal cycle where inertia terms are dominant, and it suppresses data points where small errors in measurements could have a large effect on the optimization. To achieve this, $\mathbf{w}_a$ needs to adapt to the target, so $U_{IF}[n]$ is used to determine it: During the opening phase of each pulse, $w_a[n] = 1$ if $U_{IF}[n] \geq 0.2U_{\max}$ and zero otherwise. During the closing phase, $w_a[n] = 1$ if $U_{IF}[n] \geq 0.5U_{\max}$ and zero otherwise. This asymmetry accounts for increasing losses due to entrance/exit effects when the glottal flow is still large but the glottal gap is already decreasing. The flow model has a singularity at $f_g(t) = 0$, which occurs when $A_0(t) = A_L(t) = 0$, and it is undefined when $A_0(t) = 0$ or $A_L(0) = 0$. Hence the parameter optimization is sensitive to the timing of these events and to signal values in their neighborhoods. To reduce the problems caused by this sensitivity, $w_a[n] = 0$ when A is below 10% of its maximum value regardless of the value of U_{IF}.

The values for the lossless model parameters β_1 and β_2 are solved from the NLLS problem with residual

$$\mathbf{r}_a(\beta_1, \beta_2) = \mathbf{w}_a(\mathbf{y}_a - \mathbf{f}_a) \tag{15}$$

using the trust-region-reflective algorithm (inbuilt in MATLAB's Global optimization Toolbox). The β parameters are restricted to be non-negative in the optimization, and multiple starting points are used to avoid local minima.

In Step 2, the parameters of the full model are optimized as shown in Fig. 3 (b). An NLLS problem is then set up with

$$y_{b,n} = U_{IF}[n] \quad \text{and} \quad f_{b,n} = U(A_0[n], A_L[n], \boldsymbol{\beta}), \tag{16}$$

and residual

$$\mathbf{r}_b(\boldsymbol{\beta}) = \mathbf{w}_b(\mathbf{y}_b - \mathbf{f}_b). \tag{17}$$

This involves solving the glottal flow $U(A_0[n], A_L[n], \boldsymbol{\beta})$ from Eq. (13) numerically using the explicit Euler method with $A_0[n]$ and $A_L[n]$ upsampled for the integration by a factor of 10 using linear interpolation.

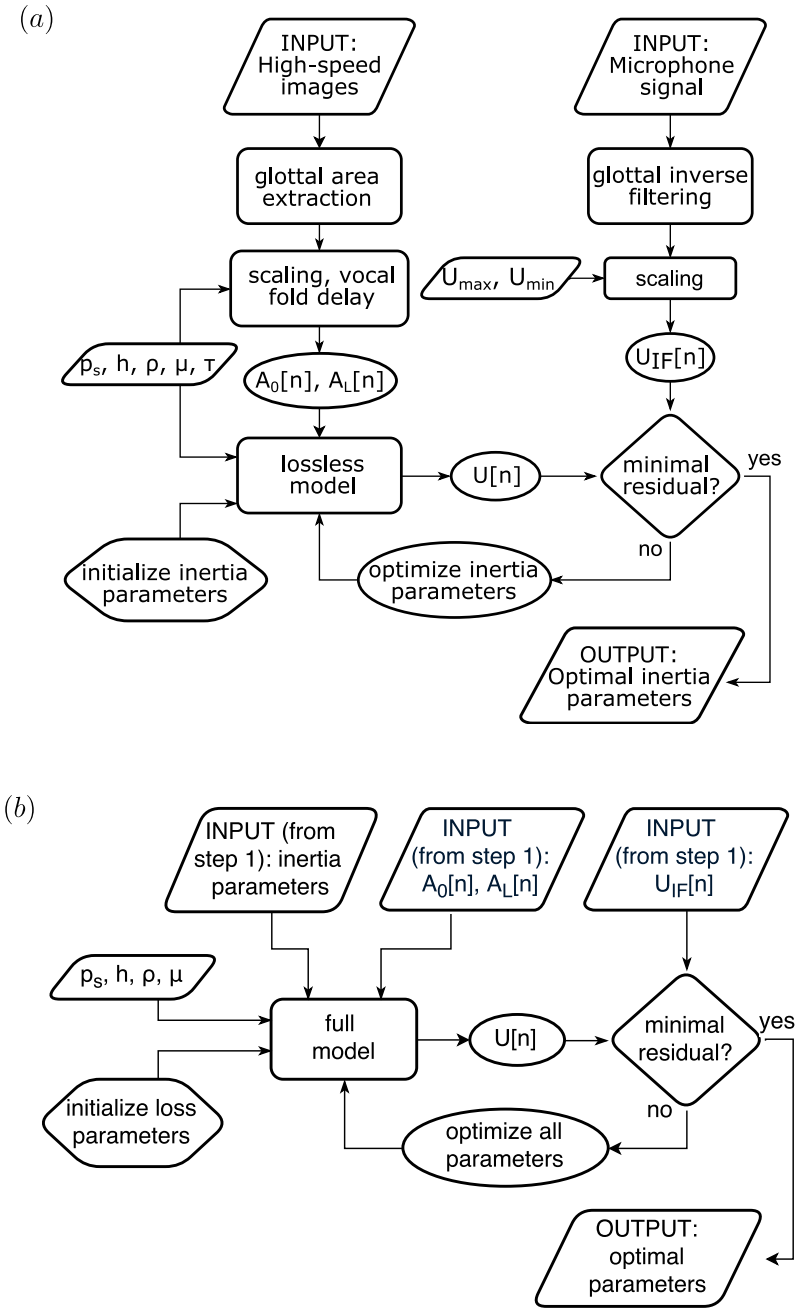

Fig. 3. Optimization of the β parameters takes place in two phases: Inertia parameters for the lossless model (β_1, β_2) are determined in Step 1 (a). Optimal parameters for the full model are determined in Step 2 (b). Inputs to Step 2 are obtained from Step 1.

Table 4
Predefined physical and physiological parameters.

Parameter	Value
subglottal pressure, p_s	1000 Pa
vocal fold length, h	18.0 mm (male) 10.0 mm (female)
maximum glottal flow, U_{max}	0.002 m^3/s
minimum glottal flow, U_{min}	0.000 m^3/s
density of air, ρ	1.12 kg/m^3
dynamic viscosity of air, μ	$18.3 \cdot 10^{-6}$ Ns/m^2

Using explicit Euler for solving Eq. (13) combines the airway parameters to one lumped parameter $\beta_1^* = \beta_1 - \Delta t \widetilde{\beta}_1$, where Δt is the time step, so that β parameter set to be optimised is $\boldsymbol{\beta} = (\beta_1^*, \beta_2, \beta_3, \beta_4)$.

The second weight function $\mathbf{w}_b$ is shown in Fig. 4. It selects the data

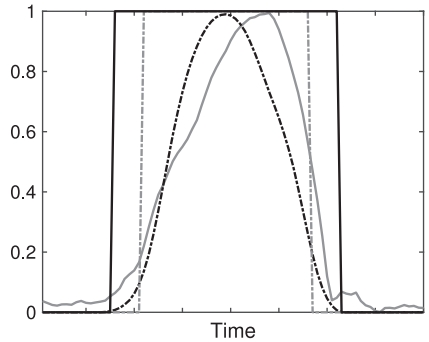

Fig. 4. Weight functions w_a (dashed gray) and w_b (solid black), defined relative to glottal flow obtained by inverse filtering U_{IF} (solid gray) and glottal area A (dashed black) pulses.

points where the glottis is open, $A(t) \geq \epsilon$, where the threshold value was set to $\epsilon = h \cdot 0.0015$ mm by trial and error. Although $\mathbf{w}_b$ also includes parts of the glottal cycle where losses are expected to be negligible, this produces the best match between model prediction and measurements. This is likely due to the difficulty finding the exact timing of loss-dominated regions over several glottal cycles.

The parameters $\boldsymbol{\beta}$ are solved again using the trust-region-reflective algorithm in MATLAB. All β parameters are restricted to non-negative values on physical grounds. The values of β_1 and β_2 obtained from Step 1 are used as a starting point for β_1^* and β_2, respectively, in Step 2. The full model optimization problem has multiple local minima, and without computing this starting point first, the optimization algorithm tends to run into convergence problems or to converge to clearly unrealistic parameter values. Allowing β_1 and β_2 to change in this step is necessary because Step 1 tends to overfit them to the data in the parts of the glottal cycle where inertial terms are large but losses are not negligible.

The value for τ is obtained using binary search over $0 \leq \tau \leq T_c$, where T_c is the duration of the closed phase. At each stage of the binary search, the β parameters were optimised using the two step approach and the residual (17) was used as the objective function.

The optimization process is summarized by the following pseudo-code.

I. Select τ from the range [0, T_c] using binary search criteria
   1. Lossless model
      (i) Set $\mathbf{y}_a$ and $\mathbf{f}_a$ using Eq. (14).
      (ii) Compute $\mathbf{w}_a$.
      (iii) Solve τ, β_1, and β_2 from NLLS problem with residual (15) subject to $\beta_1, \beta_2 \geq 0$.
   2. Full model
      (i) Initialize $\boldsymbol{\beta}$.
      (ii) Set $\mathbf{y}_b$ and $\mathbf{f}_b$ using Eq. (16).
      (iii) Compute $\mathbf{w}_b$.
      (iv) Solve $\boldsymbol{\beta}$ from NLLS problem with residual (17) subject to $\boldsymbol{\beta} \geq 0$.
II. Compare residual (17) at current and previous τ values, return to I. to update τ if not converged

This three stage parameter optimization is required to structure the process so that different parts of the target and data signals are used in physically appropriate ways. Moreover, efficiency and practicality speak against attempting to optimize all parameters at once.

5. Results

Before presenting the results of combining the physical model with the HSV and GIF data, observations are made about the data. The results of combining the three methodologies are then described in two parts: the optimal parameter values are presented first, followed by a description of the match between modeled and measured glottal flow.

5.1. Glottal area and flow

Figs. 5 and 6 show samples of the glottal area and flow data. A few pulses of the temporal signals are displayed on one panel of each figure, and the Lissajous plots (flow versus area) of the full signals are shown on the other. These figures illustrate the broad range of flow skewing visible in the data: for m05 (Fig. 5), the flow and area deviate only slightly, whereas for f02 (Fig. 6), the difference is notable. The figures are generally in line with the observations in the study by Granqvist et al. (2003), particularly with Figs. 5–9 in their investigation. Their data was obtained using an experimental setup that differs slightly from the one used in the present study, but their examples show a large range of flow–area relationships for one male and one female speaker.

Two of the samples, m01 (shown in Fig. 7) and f03, were observed to exhibit unrealistic flow skewing, where maximal flow, estimated using GIF, occurs when the vocal folds are nearly closed. This was taken to indicate a problem with the data (e.g, synchronization or stability of phonation, as discussed in Section 6); and for the rest of this work, these two samples are treated as outliers.

To quantify the variability in the flow–area relationship, two time parameters were computed: quasi-open quotient (QOQ) and quasi-speed quotient (QSQ). Previous studies have used quasi-quotients instead of the classical ones because formant ripples and noise can make the determination of the exact opening and closing instants from U_{IF} difficult (Dromey et al., 1992; Sapienza et al., 1998). Extraction of these instants from A is more reliable, but since obtaining comparable quotients for A and U_{IF} is of interest in this study, QOQ and QSQ are used. QOQ is here defined to be the proportion of each pulse duration where the signal (A or U_{IF}) is at least 25% of its maximum. QSQ is the ratio of the opening duration (i.e., when the signal is increasing from the 25% level to the pulse peak) divided by the closing duration (i.e., when the signal is decreasing from the pulse peak to the 25% level). The computed values are shown in Fig. 8.

A linear model fitted to the QOQ data, excluding m01 and f03, (using `fitlm` in MATLAB) yields a relationship

$$\mathrm{QOQ}_A = 2.40 \cdot \mathrm{QOQ}_{UIF} - 0.69 \tag{18}$$

for males and

$$\mathrm{QOQ}_A = 2.40 \cdot \mathrm{QOQ}_{UIF} - 0.75 \tag{19}$$

for females. These regressions are also shown in Fig. 8 (a). For the linear model, $R^2 = 0.839$, and p-values (t-test) for the intercept, slope, and gender effects are $p < .02$, $p < .002$, and $p < .10$, respectively. Including gender as an explanatory variable in the linear model improves the fit of the model moderately (from $R^2 = 0.736$) and causes an increase in the slope and a decrease in the intercept. This linear model suggests that (i) QOQ of the glottal area is lower than that of the glottal flow when the QOQ values are within the range usually observed (9 out of 11 samples in the data used here), but (ii) the difference is larger at low QOQ values than at high values.

These observations are broadly in line with open quotient (OQ) pairs available in the literature (directly comparable QOQ values could not be found): at low OQ values, $\mathrm{OQ}_A < \mathrm{OQ}_{UIF}$ (Krishnamurthy and Childers, 1981), whereas at high OQ values, $\mathrm{OQ}_A > \mathrm{OQ}_{UIF}$ (non-pathological speaker of Berouti et al. (1977)), though the latter reported that in most of their data the OQ_A is smaller. Pulakka (2005) also reported OQ values for synchronized A and U_{IF} signals, but comparison is difficult as their OQ_A values often fall somewhere between the two different OQ values they report for U_{IF}.

The relationship between the QSQ values of U_{IF} and A (Fig. 8 (b)) is more complicated, and a linear model is unable to explain the majority of this relationship even when the explanatory variables include gender, fundamental frequency, and speaker. Spearman's rank correlation coefficient between the two QSQ value sets is 0.627, and this correlation is significant (t-test: $p < .05$), but no significant rank correlation was found between QSQ_{UIF} and gender, fundamental frequency, or QOQ values. More qualitatively, QSQ_A is smaller than QSQ_{UIF} for all samples except m03, indicating that the flow pulses are more skewed to the right than the area pulses. Furthermore, most of the samples show area pulses that are symmetric or skewed to the left ($\mathrm{QSQ}_A \leq 1$), whereas the majority of the flow pulses are skewed to the right ($\mathrm{QSQ}_{UIF} \geq 1$).

It is worth noting that synchronization errors would in principle affect Figs. 5–7, but they would not have an impact on the time quotient values. Errors in A, however, lead to errors in the extracted time instants, which can have a large impact on the parameter values. Errors of 0.1 ms in the time instants could, for example, lead to an error of up to 0.03 in the QOQ values and 1.2 in the QSQ values.

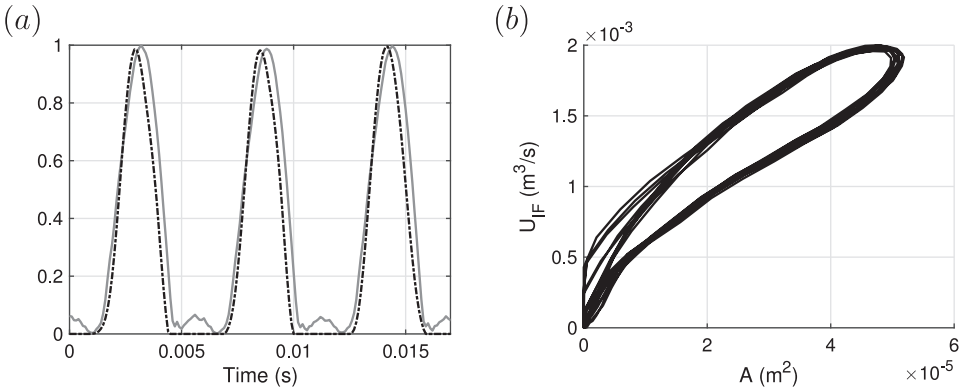

Fig. 5. Normalized glottal flow U_{IF} estimated by GIF (solid gray) and glottal area A from HSV (dashed black) for m05 (a). U_{IF} versus A for the same data (b).

5.2. Parameter optimization

The optimized parameter values of the full model are shown in Tables 5–7. These values have been computed using the constants in Table 4. This means that C_a, L_i, and L_v are scaled by $p_s/U_{\max}$, while k_t is scaled by $(p_s/U_{\max})^2$, and L_v is further scaled by h^2. Note that there is speaker and phonation dependent variation in both $p_s/U_{\max}$ and h, which is not available from the data. Emphasis is hence placed on the proportions of the parameter values. The main measure of success of the optimization is the achieved glottal pulse waveform match, as shown in Section 5.3. The optimal parameters for the lossless model are very similar to those listed in Tables 5–7; the second step of the optimization typically causes only a small change in C_a and a moderate decrease in L_i.

The value of C_a is computed by assuming $\beta_1^* = \beta_1$ since it was observed that changing the time discretization in the explicit Euler method had very little impact on the value of β_1^*. This is expected, as the computation of C_a and C_b from MRI data of vocal tracts using formulae from Aalto (2009, pp. 9–10) indicates that $\Delta t C_b \ll C_a$ at the time discretizations used. For comparison, C_{VT} calculated from MRI data of production of [œ] is in the vicinity of 1700–1800 kg/m^4 for the male who produced samples m03 and m04 and 1200–1500 kg/m^4 for the female who produced f01. A simple exponential horn model for the lower airways (Murtola, 2014) has $C_{SGT} \approx 1000$ kg/m^4. For most of the samples and gap geometries, a large proportion of the air column hence appears to act as an inertial load. Note, however, that there is a trend toward lower C_a values in linear and planar geometries, and C_a is notably small for m05 in linear and planar, and for m03 in planar gap geometries.

For male speakers, the inertial thickness L_i is mostly large compared to realistic vocal fold dimensions, whereas more reasonable values have been obtained for female speakers. The results do not provide conclusive evidence that any gap geometry matches the data better in this respect than the others. The value of L_i is also particularly sensitive to the quality and synchronization of the data. This is visible in the very low values for m01 and f03 in all geometries, and these two samples were already flagged as outliers based on their Lissajous plots. In addition, a numerical experiment using the constant gap geometry showed that allowing U_{IF} to shift relative to A by up to $\pm$ 6 time steps to simulate errors in the synchronization has a large impact on the optimal L_i value, and shifts no larger than $\pm$ 0.2 ms are able to reduce the highest L_i values to below 10 mm.

The viscous thickness L_v is smaller than L_i in the constant gap geometry as expected since the orientation of the vocal folds is typically different during the parts of the cycle where inertial and viscous terms dominate. In a few samples, the difference decreases or L_v even becomes larger than L_i when $\tau > 0$. However, in linear and planar gap geometries the majority of the L_v values optimize to approximately zero indicating that the combination of the gap geometries and the form of the viscous term did not fit the data. The fixed vocal fold length h affects the value of L_v, but a 10% change in h, for example, leads to an approximately 20% change in L_v. Hence, the fixed h cannot explain the near-zero values.

The particularly high L_i and L_v values for m03 can be explained by looking at the match between the U_{IF} and modeled flows (Fig. 9 (a)). The target flow is particularly slow to start and end at each glottal cycle. These are difficult features for the physical model to match, and the linear and planar gap geometries show classical features of overfitting: improved match at increasingly unreasonable parameter values.

In contrast to L_v, the entrance/exit effect coefficient k_t tends to increase from approximately zero in all but three samples for constant gap geometry to the order of magnitude expected based on commonly used values around 1 (e.g., van den Berg et al., 1957; Ishizaka and Flanagan, 1972) in most samples for linear and planar gap geometries. There appears to be some overlap in the two loss terms. The results indicate that either viscous losses or entrance/exit effects dominate in the optimization and, further, the optimisation appears to favor combining the viscous model with constant gap geometries and the exit effects with linear and planar gap geometries.

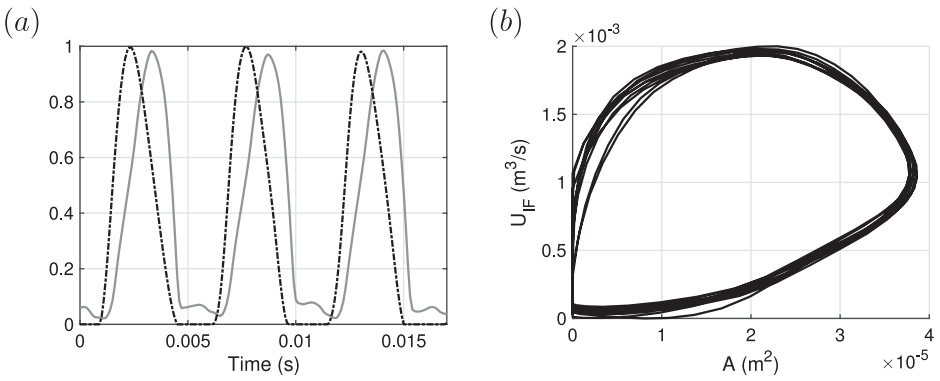

Fig. 6. Normalized glottal flow U_{IF} estimated by GIF (solid gray) and glottal area A from HSV (dashed black) for f02 (a). U_{IF} versus A for the same data (b).

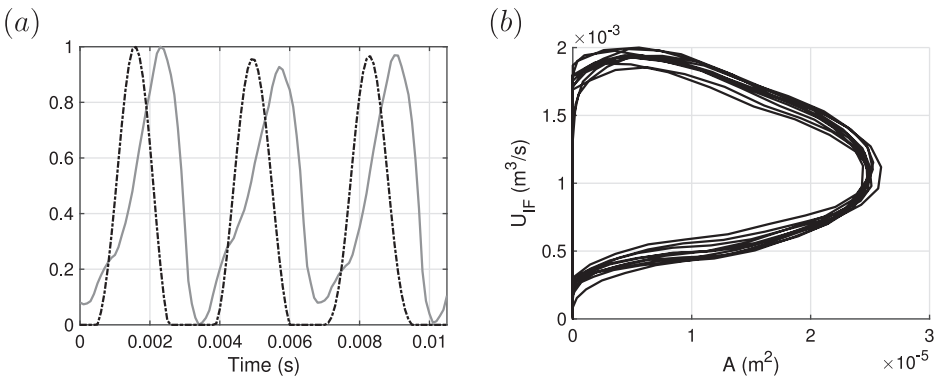

Fig. 7. Normalized glottal flow U_{IF} estimated by GIF (solid gray) and glottal area A from HSV (dashed black) for m01 (a). U_{IF} versus A for the same data (b).

A non-zero delay τ was observed to decrease the error residual for all samples except m04. For the linear gap geometry, the average optimal τ= 0.52 ms for females and 0.75 ms for males. For the planar gap geometry, this difference disappears and the averages are 0.74 ms and 0.71 ms for females and males, respectively.

5.3. Modeled glottal flow

Figs. 9 and 10 depict the glottal flow predicted by the full model with different gap geometries using the optimal parameter values as well as the target U_{IF}. For six of the samples, the changing the glottal gap geometry from the worst to the best reduced the residual by less than 40% (Fig. 11). Fig. 9 (b) shows the extreme case of m04 where all three geometries produce, at the optimum, visually identical flow signals. In the rest of the samples, changes in the residual are larger but do not necessarily produce correspondingly large visual changes in the matches (Fig. 10 (a)). In these cases the improvement is achieved by a better match at the opening and closing phases while errors at the flow peak are less affected. In all gap geometries, the match is the least satisfactory when the glottal gap is small.

Despite its very rudimentary nature, the lossless model is capable of capturing the majority of the skewing of the glottal flow compared to the glottal area, but the full model produces overall slightly better matches. The full model in particular is better able to match the pulse peak height, and it also improves the general match during the opening phase. Even the full model, however, cannot match the slowing in the rate of decrease immediately before closure, which is often evident in the flow estimated by GIF.

Any flow present when the vocal folds are closed is, by the definition of Eq. (2), outside the scope of the model. This is visible in all the matched pulses but particularly in the optimized model flow for samples m01 and f03 (the latter is shown in Fig. 10 (b)). Since A and U_{IF} for these samples do not appear to match, the model ignores the area signal and produces triangular pulses when the vocal folds are open.

The values for time parameters QOQ and QSQ for U_{IF} and U produced by the full model with planar gap geometry are shown in Fig. 12. For this figure, U_{IF} is allowed to shift up to $\pm$ 0.6 ms relative to A to account for possible synchronization errors. The relationship between the quasi-parameters can be well predicted using linear models (excluding m01 and f03). For QOQ,

$$\mathrm{QOQ}_U = 2.12 \cdot \mathrm{QOQ}_{UIF} - 0.54 \tag{20}$$

with $R^2 = 0.936$ and p-values below .001. The similarity between Eq. (20) and Eqs. (18)–(19) is notable. This is because the opening and closing instants of the modeled flow are determined by the area signal, and only the use of QOQ instead of the classical OQ enables the equations to differ. The higher the cut-off level is set in determining QOQ, the more the skewing of the flow contributes to its value. Unlike in the case of the relationship between QOQ_A and QOQ_{UIF}, adding gender as an explanatory variable does not noticeably improve the fit of Eq. (20).

For *QSQ*, gender is a significant factor, so that for males

$$\mathrm{QSQ}_U = 1.89 \cdot \mathrm{QSQ}_{UIF} - 0.87 \tag{21}$$

and for females

$$\mathrm{QSQ}_A = 1.89 \cdot \mathrm{QOQ}_{UIF} - 0.26. \tag{22}$$

For this fit, $R^2 = 0.968$, and p-values for the intercept, slope, and gender effect are $p < .012$, $p < .00002$, and $p < .0028$, respectively. The QSQ for the modeled flow is hence higher than for the flow obtained by GIF, except when QSQ is very low, but the correlation between the two is high. Note that the high value of the slope in Eqs. (21) and (22) is at least partially explained by the way the optimization is carried out. Matching is done over full pulse waveforms, which appears to result in an increasing overestimation of the skewing, as QSQ increases when this skewing is measured using only the three time instants determining QSQ.

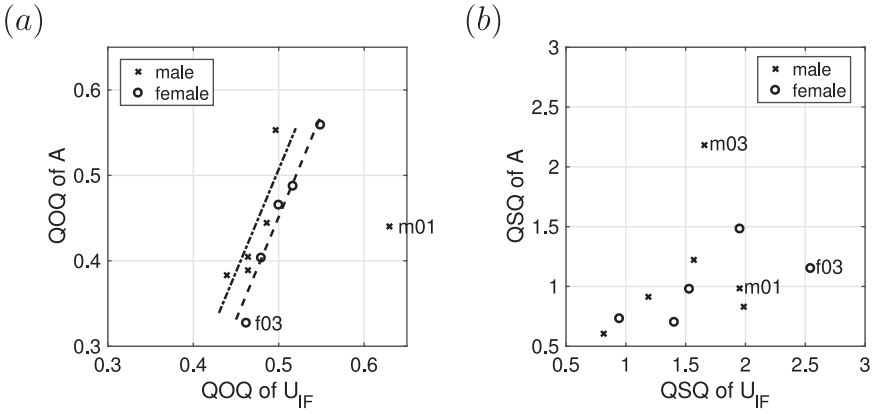

Fig. 8. Time parameter values, quasi-open quotient (a) and quasi-speed quotient (b), for glottal area A versus the same parameters for glottal flow U_{IF} estimated by GIF.

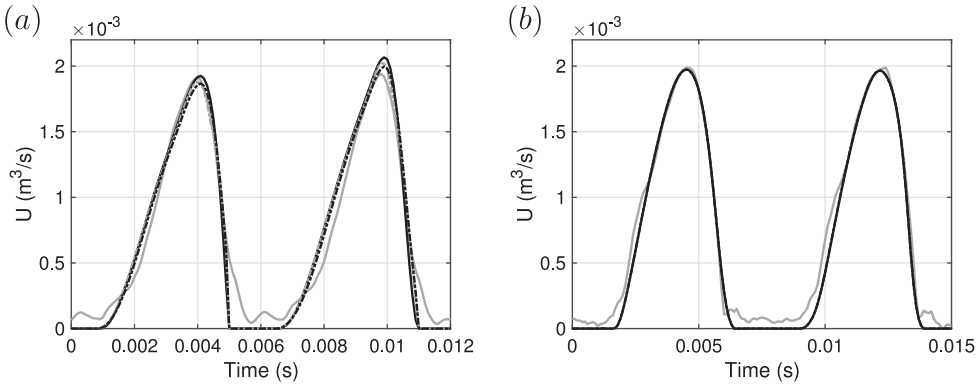

Fig. 9. Glottal flow obtained by inverse filtering U_{IF} (solid gray), and flows predicted by the full model with constant gap geometry (solid black), linear gap geometry (dashed gray) and planar gap geometry (dashed black) for m03 (a) and m04 (b).

Table 5
Optimized model parameter values for the full model in constant gap geometry. The outliers indicated by their Lissajous plots are marked with an asterisk.

Male				
Sample	C_a (kg/m^4)	L_i (mm)	L_v (mm)	k_t
m01*	983	0.054	0.022	0.001
m02	1700	11.032	0.157	0.006
m03	844	24.271	1.376	0.000
m04	734	5.355	0.000	1.051
m05	349	6.157	0.000	0.716
m06	547	11.450	0.000	2.391
Female				
Sample	C_a (kg/m^4)	L_i (mm)	L_v (mm)	k_t
f01	526	3.094	0.344	0.000
f02	1175	3.063	0.085	0.000
f03*	787	0.003	0.000	0.002
f04	1108	3.191	0.112	0.000
f05	1980	1.790	0.150	0.000

Table 6
Optimized model parameter values for the full model in linear gap geometry. The outliers indicated by their Lissajous plots are marked with an asterisk.

Male					
Sample	C_a (kg/m^4)	L_i (mm)	L_v (mm)	k_t	τ (ms)
m01*	830	0.087	0.220	0.136	1.0
m02	538	1.535	0.004	0.634	1.1
m03	450	35.136	9.122	0.000	0.8
m04	734	5.355	0.000	1.051	0.0
m05	45	2.737	0.008	1.143	0.6
m06	191	10.282	2.616	3.629	1.0
Female					
Sample	C_a (kg/m^4)	L_i (mm)	L_v (mm)	k_t	τ (ms)
f01	472	3.429	1.547	0.002	0.3
f02	1016	0.101	0.000	0.311	0.7
f03*	641	0.077	0.000	0.015	0.4
f04	480	0.514	0.002	0.172	0.7
f05	1106	1.229	0.000	0.489	1.0

6. Discussion

In this study, three methods of investigating speech production have been combined. A glottal area signal has been extracted from HSV data, glottal airflow has been estimated from a free-field microphone signal using GIF, and a physical model of the glottal flow has been optimized to match these data. Time-based quasi-quotients have been used to present quantitative data on the relationships between the HSV area signals and the flows estimated by GIF, as well as between flows obtained from the physical model and GIF.

A simple linear relationship (Eqs. (18) and (19)) is able to explain the majority of the difference between the QOQ values of HSV area and GIF flow. The same does not appear to be possible for the QSQ which describes the symmetry of the pulses. The known variables in the data (gender, individual, fundamental frequency) are hence not sufficient to determine the relationship between the area and flow.

The study by Pulakka (2005) suggested that the relationship between the time quotients of area and flow signals can vary both with phonation type and with intensity of the utterance. While data with incomplete glottal closure was not used for the current investigation, the type of phonation was not otherwise controlled. The samples may hence contain any phonation type from pressed to nearly breathy. Likewise, intensity of the utterances was not controlled. The measurements were challenging for most of the test subjects, and additional requirements and repetitions would have increased the strain on the test subjects in some cases to an unbearable level.

The two data samples, m01 and f03, were likely observed to be outliers in their Lissajous plot behavior due to a combination of factors. The limited accuracy of the synchronisation of the different signals may contribute to the problem but synchronization alone cannot explain the phenomenon fully. Both samples were utterances at high fundamental frequency. While high fundamental frequency, *per se*, does not hinder conducting a GIF analysis, it typically causes impairment of the estimated glottal flow due to biasing of the estimated resonances of the vocal tract by sparse harmonics (Alku, 2011; Drugman et al., 2014). The glottal area signal and the flow estimated by GIF may also contain noise and synchronization errors. When the fundamental frequency is high, the impact of these errors in the parameter optimization can become large, as there are fewer data points per glottal pulse due to the fixed sampling rate. The short duration of the glottal cycle can also make unusual vocal fold behavior difficult to detect. The speaker producing m01 and m02, in particular, was noted to be prone to asymmetric vocal fold oscillations in other samples of the larger dataset although no such behavior was seen in the two video samples used in this study. It is worth noting, however, that there were no problems with the data or the parameter optimization of the third high pitch sample f01.

Using the linear or planar glottal gap geometries improved the match between the target flow U_{IF} and the modeled flow U compared to the constant gap geometry. However, the more realistic gap geometries did not produce overall more realistic parameter values. In fact, there is some indication that the more complicated models may be overfitting the data.

The parameter optimisation appeared to favor combining the viscous loss model with the constant gap geometry and entrance/exit effects models with linear and planar gaps geometries. Both terms were notably larger than zero in only a few cases, and there were also a few

Table 7
Optimized model parameter values for the full model in planar gap geometry. The outliers indicated by their Lissajous plots are marked with an asterisk.

Male					
Sample	C_a (kg/m^4)	L_i (mm)	L_v (mm)	k_t	τ (ms)
m01*	873	0.029	0.221	0.110	0.9
m02	573	0.550	0.000	0.642	1.0
m03	5	55.802	5.313	0.000	1.0
m04	734	5.910	0.000	1.042	0.0
m05	43	2.289	0.008	1.183	0.6
m06	259	5.175	0.039	3.680	0.8
Female					
Sample	C_a (kg/m^4)	L_i (mm)	L_v (mm)	k_t	τ (ms)
f01	334	6.225	9.122	0.005	0.8
f02	1016	0.107	0.000	0.311	0.7
f03*	555	0.088	0.000	0.027	0.5
f04	470	0.337	0.007	0.178	0.7
f05	1112	1.594	0.005	0.484	1.0

cases where both terms optimised to approximately zero. There are a number of possible factors why the loss terms did not fit the data.

First, the loss terms are particularly sensitive to the glottal geometry used. The rectangular cross sectional glottal areas used in the constant and linear gap geometries, as well as the integrated estimate for the viscous loss in the planar gap geometry, all tend to underestimate L_v. In contrast, assuming that the vocal folds are open for their entire length h (as all three geometries do) leads to overestimation of L_v when the gap is in reality shorter, e.g, at opening or closing. A partial solution to the latter problem is to extract the length of the gap in the z-direction and scale it with the fixed vocal fold length to obtain $\widetilde{h}(t)$. This data can be used in Eq. (7) instead of h for the constant gap geometry. The results for this gap geometry variation were not reported here because this lead neither to noticeably decreased average matching errors nor to overall more realistic parameter values when compared to the constant gap geometry. It is possible that a combination of $\widetilde{h}(t)$ with linear or planar geometries could lead to improvements of the model performance. However, estimation of $\widetilde{h}(t, x)$ at $x = 0$ and $x = L$ is challenging due to noise in the signal extracted from HSV.

Second, the use of the inverse cubic power law of the Poiseuille formula and constant k_t have been questioned even in simplified glottal geometries (Fulcher and Scherer, 2011; Fulcher et al., 2013). Fulcher et al. (2013) suggested using $g^{-2.59}$ in the viscous term, and Fulcher and Scherer (2011) added a term proportional to U^2/g^3 to improve the entrance/exit model when g is small. Both of these changes were suggested based on model experiments, so their direct application to speech data may not result in the desired improvements. They do, however, suggest that adjusting the power of g in the loss terms may lead to a more favorable match between model and data.

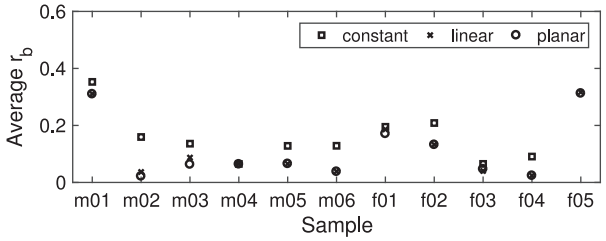

Fig. 11. Average residue r_b (i.e., error in match between U_{IF} and modeled flow) in different glottal gap geometries.

Third, the intervals where the glottal losses are largest are also the intervals where the data are most inaccurate. The instants of opening and closure can be difficult to determine, and any inaccuracies in the extraction of the glottal area in the few frames immediately following the opening instant or before closure will cause interpolation errors. GIF has also a tendency to show the poorest estimation accuracy in the vicinity of critical time instants (i.e., opening and closure) of the glottal cycle compared to other parts of the cycle.

Finally, it could simply be that the data is unsuitable for parameter estimation near closure and opening instants. For example, Hertegård and Gauffin (1995) and Granqvist et al. (2003) suggested that the non-zero flow when the vocal folds are closed (which is often seen in flow estimated by GIF) could be caused by the vertical movement of the vocal folds or the displacement of air due to the mucosal wave. That kind of behavior would be present in the flow data but not in the area function or the model, making good matching difficult.

The dataset used in this investigation was limited and hence inference of causes behind observed phenomena is difficult. Unfortunately, acquisition of additional data is not straightforward for several reasons: It was observed that not every volunteer was able to successfully carry out the phonation tasks. Even when the speakers succeeded, the tasks were strenuous and uncomfortable making higher number of repetitions infeasible. Checking the quality of the data fully during measurement sessions was also not feasible due to time considerations, and hence not every utterance recorded contained a sample where all three signals were simultaneously of sufficiently high quality for investigations such as this. Additional selection criteria, such as full glottal closure, further decrease the amount of available data.

One way to increase the size of the used dataset is to include samples where the glottis remain partially open throughout phonation. The physical model (2) is equally valid for phonation where there is no full glottal closure but the lossless model (3) used for parameter initialisation does require it. Hence, one of the challenges is developing optimisation procedures that produce a physiologically feasible optimum. The other major challenge lies in interpretation of the results. The DC component of the glottal flow is another unknown parameter that will either need to be arbitrarily fixed, requiring sensitivity analysis for the results, or optimized, requiring optimization and interpretation of

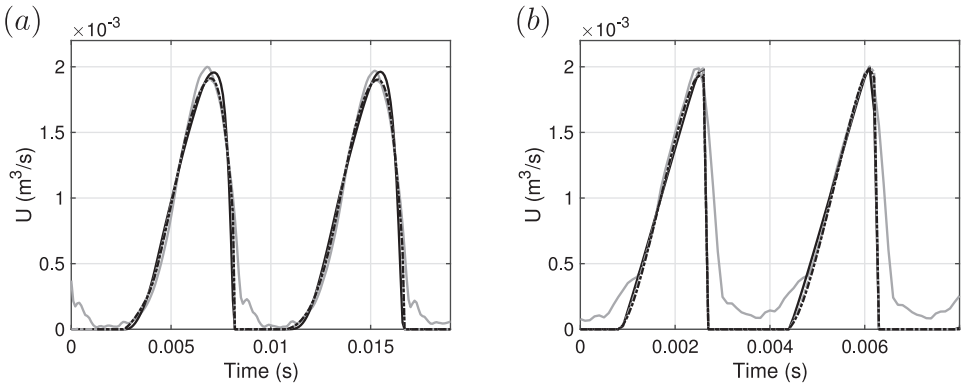

Fig. 10. Glottal flow obtained by inverse filtering U_{IF} (solid gray), and flows predicted by the full model with constant gap geometry (solid black), linear gap geometry (dashed gray) and planar gap geometry (dashed black) for m02 (a) and f03 (b).

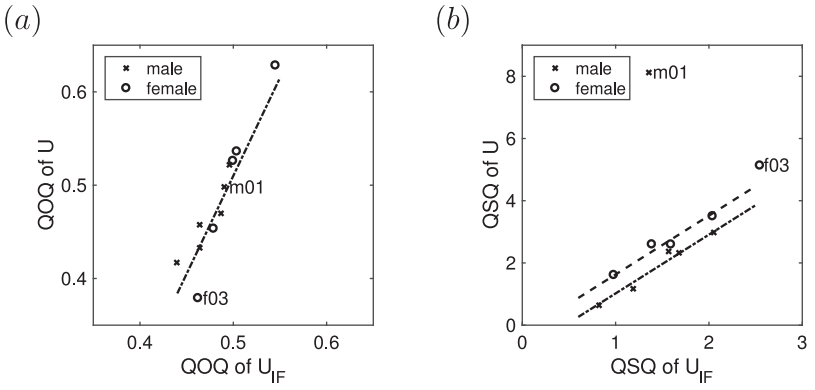

Fig. 12. Quasi-open quotient (a) and quasi-speed quotient (b) computed from flows obtained by GIF (x-axis) and the full model (y-axis). The data has been corrected for synchronization.

mixtures of unknown parameters.

The setup preliminarily experimented with in this study can in principle be used clinically with no additional discomfort for patients because in addition to HSV (which is widely used in voice clinics today) only synchronized recording of the acoustical speech pressure wave is needed. Thus, in addition to conducting HSV studies in patients, the clinician can also take advantage of the glottal flow through GIF and utilize the glottal flow model fitting procedure presented in this study. The use of the setup can also be expanded to include voice disorders enabling comparison of model parameter values between normal and disordered voices.

7. Conclusion

By combining three methodologies (HSV, GIF, and physical modeling), a physical model of the glottal flow has been successfully parameterized to match data obtained from natural speech production. Despite the challenges posed by the multidisciplinary nature of this problem, i.e., finding common ground for mathematical modeling, signal processing, and *in vivo* human experiments, the results are encouraging: The physical model is, to a large extent, capable of capturing the salient features of a natural glottal flow waveform. This opens up the possibility of constructing physically motivated models where the glottal flow is controlled via the area. This is an appealing prospect since the glottal area is expected to be easier to parameterize than the flow.

The optimized values of the physical model parameters indicate that the skewing of the glottal flow pulses can, in this model, be largely explained by the vocal and subglottal tracts acting as inertial loads. The inertia within the glottis emerged as a feature that is very sensitive to data quality and synchronization, and hence it can be seen as a good candidate for automatic data checking. The results regarding the glottal loss parameters were inconclusive. Based on these results, it is, however, likely that not all combinations of model glottal geometries and the formulae used for viscous losses and entrance/exit effects are suitable for modeling natural speech.

The skewing of the glottal flow compared to the glottal area is a well-documented phenomenon, and the results of this study provide further valuable information about it. It was observed that the QOQ values of the glottal flow and area exhibit an easily quantifiable dependence, while the QSQ values in this flow–area relationship can only be noted to be significantly correlated. Furthermore, gender appears to play a role in the difference between the glottal flow and area QOQ values, but no significant correlation could be seen between QSQ differences and gender, fundamental frequency, or QOQ.

Acknowledgments

This study was funded by the Academy of Finland (projects no. 284671 and 312490). The third author has received support from the Magnus Ehrnrooth Foundation. The authors would like to thank Prof. Erkki Vilkman and Dos. Elina Isotalo for their support and contributions to the success of this research.

References

Aalto, A., 2009. A Low-Order Glottis Model with Nonturbulent Flow and Mechanically Coupled Acoustic Load. Helsinki University of Technology, Department of Mathematics and Systems Analysis Master's thesis.

Aalto, D., Aaltonen, O., Happonen, R.-P., Jääsaari, P., Kivelä, A., Kuortti, J., Luukinen, J.-M., Malinen, J., Murtola, T., Parkkola, R., Saunavaara, J., Soukka, T., Vainio, M., 2014. Large scale data acquisition of simultaneous MRI and speech. Appl. Acoust. 83, 64–75. http://dx.doi.org/10.1016/j.apacoust.2014.03.003.

Airaksinen, M., Raitio, T., Story, B., Alku, P., 2014. Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. IEEE/ACM Trans. Audio SpeechLang. Process. 22 (3), 596–607. http://dx.doi.org/10.1109/TASLP.2013.2294585.

Alku, P., 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Commun. 11 (2), 109–118. http://dx.doi.org/10.1016/0167-6393(92)90005-R.

Alku, P., 2011. Glottal inverse filtering analysis of human voice production - a review of estimation and parameterization methods of the glottal excitation and their applications. Sadhana 36 (5), 623–650. http://dx.doi.org/10.1007/s12046-011-0041-5.

Alku, P., Horáček, J., Airas, M., Griffond-Boitier, F., Laukkanen, A.-M., 2006. Performance of glottal inverse filtering as tested by aeroelastic modelling of phonation and FE modelling of vocal tract. Acta Acust.United Acust. 92, 717–724.

Alku, P., Pohjalainen, H., Airaksinen, M., 2017. Aalto Aparat - a freely available tool for glottal inverse filtering and voice source parameterization. Subsidia: Tools and Resources for Speech Sciences.

Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A.-M., Story, B.H., 2013. Formant frequency estimation of high-pitched vowels using weighted linear prediction. J. Acoust. Soc. Am. 134 (2), 1295–1313. http://dx.doi.org/10.1121/1.4812756.

van den Berg, J., Zantema, J.T., Doornenbal, P., 1957. On the air resistance and the Bernoulli effect of the human larynx. J. Acoust. Soc. Am. 29 (5), 626–631. http://dx.doi.org/10.1121/1.1908987.

Berouti, M., Childers, D., Paige, A., 1977. Glottal area versus glottal volume-velocity. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '77. 2. pp. 33–36. http://dx.doi.org/10.1109/ICASSP.1977.1170184.

Childers, D.G., Naik, J.M., Larar, J.N., Krishnamurthy, A.K., Moore, G.P., 1985. Electroglottography, Speech, and Ultra-high Speed Cinematography. In: Titze, I.R., Scherer, R.C. (Eds.), Vocal Fold Physiology. The Dencer Center For The Performing Arts, Denver, pp. 202–220.

Cranen, B., Boves, L., 1985. Pressure measurements during speech production using semiconductor miniature pressure transducers: impact on models for speech production. J. Acoust. Soc. Am. 77 (4), 1543–1551. http://dx.doi.org/10.1121/1.391997.

Cranen, B., Boves, L., 1985. Aerodynamic aspects of voicing: glottal pulse skewing revisited. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '85 10, 1085–1088. http://dx.doi.org/10.1109/ICASSP.1985.1168104.

Cranen, B., Boves, L., 1988. On the measurement of glottal flow. J. Acoust. Soc. Am. 84 (3), 888–900. http://dx.doi.org/10.1121/1.396658.

Döllinger, M., Hoppe, U., Hettlich, F., Lohscheller, J., Schuberth, S., Eysholdt, U., 2002. Vibration parameter extraction from endoscopic image series of the vocal folds. IEEE Trans. Biomed. Eng. 49 (8), 773–781. http://dx.doi.org/10.1109/TBME.2002.800755.

Drioli, C., 2005. A flow waveform-matched low-dimensional glottal model based on physical knowledge. J. Acoust. Soc. Am. 117 (5), 3184–3195. http://dx.doi.org/10.1121/1.1861234.

Drioli, C., Foresti, G.L., 2015. Accurate glottal model parametrization by integrating audio and high-speed endoscopic video data. Signal Image Video Process. 9 (6), 1451–1459. http://dx.doi.org/10.1007/s11760-013-0597-0.

Drioli, C., Foresti, G.L., 2015. Data-driven vocal folds models for the representation of both acoustic and high speed video data. 2015 International Joint Conference on Neural Networks (IJCNN). pp. 1–6. http://dx.doi.org/10.1109/IJCNN.2015.7280685.

Dromey, C., Stathopoulos, E.T., Sapienza, C.M., 1992. Glottal airflow and electroglottographic measures of vocal function at multiple intensities. J. Voice 6 (1), 44–54. http://dx.doi.org/10.1016/S0892-1997(05)80008-6.

Drugman, T., Alku, P., Alwan, A., Yegnanarayana, B., 2014. Glottal source processing: from analysis to applications. Comput. Speech Lang. 28 (5), 1117–1138. http://dx.doi.org/10.1016/j.csl.2014.03.003.

Drugman, T., Bozkurt, B., Dutoit, T., 2012. A comparative study of glottal source estimation techniques. Comput. Speech Lang. 26 (1), 20–34. http://dx.doi.org/10.1016/j.csl.2011.03.003.

Elie, B., Laprie, Y., 2016. Extension of the single-matrix formulation of the vocal tract: consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink. Speech Commun. 82, 85–96. http://dx.doi.org/10.1016/j.specom.2016.06.002.

Eysholdt, U., Rosanowski, F., Hoppe, U., 2003. Vocal fold vibration irregularities caused by different types of laryngeal asymmetry. Eur. Arch. Oto-Rhino-Laryngol. 260 (8), 412–417. http://dx.doi.org/10.1007/s00405-003-0606-y.

Fant, G., 1960. Acoustic Theory of Speech Production. Mouton, The Hague.

Fant, G., Liljencrants, J., Lin, Q.-G., 1985. A four-parameter model of glottal flow. STL-QPSR 1–13.

Fulcher, L.P., Scherer, R.C., 2011. Phonation threshold pressure: comparison of calculations and measurements taken with physical models of the vocal fold mucosa. J. Acoust. Soc. Am. 130 (3), 1597–1605. http://dx.doi.org/10.1121/1.3605672.

Fulcher, L.P., Scherer, R.C., Powell, T., 2011. Pressure distributions in a static physical model of the uniform glottis: entrance and exit coefficients. J. Acoust. Soc. Am. 129 (3), 1548–1553. http://dx.doi.org/10.1121/1.3514424.

Fulcher, L.P., Scherer, R.C., Powell, T., 2013. Viscous effects in a static physical model of the uniform glottis. J. Acoust. Soc. Am. 134 (2), 1253–1260. http://dx.doi.org/10.1121/1.4812859.

Gómez-Vilda, P., Fernández-Baillo, R., Nieto, A., Díaz, F., Fernández-Camacho, F., Rodellar, V., Álvarez, A., Martínez, R., 2007. Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters. J. Voice 21 (4), 450–476. http://dx.doi.org/10.1016/j.jvoice.2006.01.008.

Granqvist, S., Hertegård, S., Larsson, H., Sundberg, J., 2003. Simultaneous analysis of vocal fold vibration and transglottal airflow: exploring a new experimental setup. J. Voice 17 (3), 319–330. http://dx.doi.org/10.1067/S0892-1997(03)00070-5.

Guðnason, J., Mehta, D.D., Quatieri, T.F., 2015. Evaluation of speech inverse filtering techniques using a physiologically based synthesizer. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4245–4249. http://dx.doi.org/10.1109/ICASSP.2015.7178771.

Hertegård, S., Gauffin, J., 1995. Glottal area and vibratory patterns studied with simultaneous stroboscopy, flow glottography, and electroglottography. J. Speech Hear. Res. 38 (1), 85-100.

Ishizaka, K., Flanagan, J.L., 1972. Synthesis of voiced sounds from a two mass model of the vocal cords. Bell Syst. Tech. J. 51, 1233–1268.

Krishnamurthy, A., Childers, D., 1981. Vocal fold vibratory patterns: comparison of film and inverse filtering. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '81. 6. pp. 133–136. http://dx.doi.org/10.1109/ICASSP.1981.1171353.

Lohscheller, J., Toy, H., Rosanowski, F., Eysholdt, U., Döllinger, M., 2007. Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. Med. Image Anal. 11 (4), 400–413. http://dx.doi.org/10.1016/j.media.2007.04.005.

Lous, N.J.C., Hofmans, G.C.J., Veldhuis, R.N.J., Hirschberg, A., 1998. A symmetrical two-mass vocal-fold model coupled to vocal tract and trachea, with application to prosthesis design. Acta Acust.United Acustica 84 (6), 1135–1150.

Lucero, J.C., 1996. Relation between the phonation threshold pressure and the prephonatory glottal width in a rectangular glottis. J. Acoust. Soc. Am. 100 (4), 2551–2554. http://dx.doi.org/10.1121/1.417365.

Mehta, D.D., Zañartu, M., Quatieri, T.F., Deliyski, D.D., Hillman, R.E., 2011. Investigating acoustic correlates of human vocal fold vibratory phase asymmetry through modeling and laryngeal high-speed videoendoscopy. J. Acoust. Soc. Am. 130 (6), 3999–4009. http://dx.doi.org/10.1121/1.3658441.

Mergell, P., Herzel, H., Titze, I.R., 2000. Irregular vocal-fold vibration - high-speed observation and modeling. J. Acoust. Soc. Am. 108 (6), 2996–3002. http://dx.doi.org/10.1121/1.1314398.

Murtola, T., 2014. Modelling Vowel Production. Aalto University School of Science, Department of Mathematics and Systems Analysis Licentiate thesis.

Pelorson, X., Hirschberg, A., van Hassel, R.R., Wijnands, A.P.J., Auregan, Y., 1994. Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. application to a modified two-mass model. J. Acoust. Soc. Am. 96 (6), 3416–3431. http://dx.doi.org/10.1121/1.411449.

Pulakka, H., 2005. Analysis of Human Voice Production Using Inverse Filtering, High-speed Imaging, and Electroglottography. Helsinki University of Technology, Department of Computer Science and Engineering Master's thesis.

Qin, X., Wang, S., Wan, M., 2009. Improving reliability and accuracy of vibration parameters of vocal folds based on high-speed video and electroglottography. IEEE Trans. Biomed. Eng. 56 (6), 1744–1754. http://dx.doi.org/10.1109/TBME.2009.2015772.

Rothenberg, M., 1973. A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. J. Acoust. Soc. Am. 53 (6), 1632–1645. http://dx.doi.org/10.1121/1.1913513.

Rothenberg, M., 1981. Acoustic interaction between the glottal source and the vocal tract. In: Stevens, K.N., Hirano, M. (Eds.), Vocal Fold Physiology. Vol. 21. University of Tokyo Press, Japan, pp. 305–328.

Sapienza, C.M., Stathopoulos, E.T., Dromey, C., 1998. Approximations of open quotient and speed quotient from glottal airflow and EGG waveforms: effects of measurement criteria and sound pressure level. J. Voice 12 (1), 31–43. http://dx.doi.org/10.1016/S0892-1997(98)80073-8.

Schwarz, R., Hoppe, U., Schuster, M., Wurzbacher, T., Eysholdt, U., Lohscheller, J., 2006. Classification of unilateral vocal fold paralysis by endoscopic digital high-speed recordings and inversion of a biomechanical model. IEEE Trans. Biomed. Eng. 53 (6), 1099–1108. http://dx.doi.org/10.1109/TBME.2006.873396.

Sparrow, E.M., 1962. Laminar flow in isosceles triangular ducts. AIChE J. 8 (5), 599–603.

Steinecke, I., Herzel, H., 1995. Bifurcations in an asymmetric vocal-fold model. J. Acoust. Soc. Am. 97 (3), 1874–1884. http://dx.doi.org/10.1121/1.412061.

Story, B.H., Titze, I.R., 1995. Voice simulation with a body-cover model of the vocal folds. J. Acoust. Soc. Am. 97 (2), 1249–1260. http://dx.doi.org/10.1121/1.412234.

Titze, I.R., 1984. Parameterization of the glottal area, glottal flow, and vocal fold contact area. J. Acoust. Soc. Am. 75 (2), 570–580. http://dx.doi.org/10.1121/1.390530.

Titze, I.R., 1988. The physics of small-amplitude oscillation of the vocal folds. J. Acoust. Soc. Am. 83 (4), 1536–1552. http://dx.doi.org/10.1121/1.395910.

Titze, I.R., 2006. Voice training and therapy with a semi-occluded vocal tract: rationale and scientific underpinnings. J. Speech Lang. Hear. Res. 49 (2), 448–459. http://dx.doi.org/10.1044/1092-4388(2006/035).

Švancara, P., Horáček, J., 2006. Numerical modelling of effect of tonsillectomy on production of Czech vowels. Acta Acust.United Acust. 92, 681–688.

Wong, D., Markel, J., Gray, A., 1979. Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Trans. Acoust. 27 (4), 350–355. http://dx.doi.org/10.1109/TASSP.1979.1163260.

Wurzbacher, T., Döllinger, M., Schwarz, R., Hoppe, U., Eysholdt, U., Lohscheller, J., 2008. Spatiotemporal classification of vocal fold dynamics by a multimass model comprising time-dependent parameters. J. Acoust. Soc. Am. 123 (4), 2324–2334. http://dx.doi.org/10.1121/1.2835435.

Wurzbacher, T., Schwarz, R., Döllinger, M., Hoppe, U., Eysholdt, U., Lohscheller, J., 2006. Model-based classification of nonstationary vocal fold vibrations. J. Acoust. Soc. Am. 120 (2), 1012–1027. http://dx.doi.org/10.1121/1.2211550.

Zhang, Y., Jiang, J.J., 2008. Nonlinear dynamic mechanism of vocal tremor from voice analysis and model simulations. J. Sound Vib. 316 (1–5), 248–262. http://dx.doi.org/10.1016/j.jsv.2008.02.026.